Dataset Linking in a Multilingual Linked Open Data Context

Melkamu Beyene
Addis Ababa University
Addis Ababa, Ethiopia

Pierre-Edouard Portier
INSA de Lyon
Lyon, France

Solomon Atnafu
Addis Ababa University
Addis Ababa, Ethiopia

Sylvie Calabre
INSA de Lyon
Lyon, France

ABSTRACT

Although, the syntactical and structural heterogeneities among inter-language linked open data (LOD) data sources bring many challenges, entity co-reference resolution in a multilingual linked open data (MLOD) setting is not well studied.

In this research, a three phase approach is proposed. First, statistical relational learning (SRL) with factorization of three way tensor is used to compute structural similarity between entities. Second, textual data from the Web of documents is associated in order to increase our knowledge of entities. Through a latent Dirichlet allocation (LDA), entities' textual data is projected into a cross-lingual topic space. This cross-lingual topic space is used to find textual similarities between entities. Third, a belief aggregation strategy is used to combine the structural and textual similarity results into a global similarity score.

We have shown by experiments that our algorithm outperforms state of the art approaches based on tensor decomposition for the task of entity co-reference resolution in a MLOD setting.

Keywords

Linked data; cross lingual dataset linking; entity co-reference resolution;LDA

1. INTRODUCTION

The Linked Data publishing principles recommend data publishers to provide owl: *sameAs* links between resources referring to the same real world object or concept [1]. However, due to language barriers, data publishers didn't set such identity indicating links among inter-language LOD sources. The MLOD environment can be considered as an extreme scenario where maximum structural and syntactical heterogeneities can be observed. Various situations such as name variations, abbreviations, misspellings, inaccurate translations, differences in granularity of conceptualization across languages and cultures, difference in syntax and semantics of ontologies, and open world assumption of the MLOD may mislead entity co-reference resolution systems. In order to detect co-referent instances in the MLOD environment, it is important to use the most appropriate and relevant information and discard misleading information. To this end, our proposed system combines similarity analysis results from two types of evidences.

The first evidence is entities structural neighborhood. It is used by assuming the structural neighborhood of an entity is a useful source of information to determine its identity [2,3,4,5]. To model the structural neighborhood of entities, SRL techniques based on tensor factorizations are used since these techniques have started to bring performance improvement for various relational learning tasks [6,7,8,9,10] and due to their ability to balance expressiveness and scalability [5]. The RESCAL tensor factorization model is becoming popular for many relational learning tasks in a LOD environment including to the entity co-reference resolution task [2,11]. The RESCAL model uses the structural neighborhood of entities to decide whether two URIs are references of the same entity as it has the ability to exploit relationship correlations across entities and predicates. For instance, if two URIs share many neighbors in-common (i.e. entity or predicate), RESCAL assumes the two URIs are co-referent. However, in a highly heterogeneous environment such as in a MLOD context, RESCAL may lack this ability due to the existence of different syntactical representation among identical subjects, predicates and objects of an rdf triple even if they refer to the same entity. In this paper, we reduce heterogeneities of the MLOD environment by canonicalizing different syntactical representations that refer to the same real world object or concept into a common representation. This also increases the scalability of the model as it reduces the number of distinct entities and relationships. Here, the researchers used existing equivalent relations and reasoning (i.e. rdfs and owl) to merge syntactically distinct URI references referring to the same entity into a common representation. Thus, we have showed the fact that a better performance can be achieved by combining SRL techniques with reasoning for the task of entity co-reference detection. Moreover, RESCAL assumes all relations (i.e. predicates) are equally important. This assumption is unrealistic for the task of entity co-reference resolution as some relations are more informative than others. Therefore, one need to propose an approach that takes predicate informativeness weight into consideration. The RESCAL tensor factorization model has no mechanism to add predicates informativeness weight. In this paper, we use the Decomposition into Directional Components (DEDICOM) model and pred-

icates informativeness weight is introduced as an additional constraint [22,56].

The second evidence is entities textual descriptions from unstructured sources. We purposely exclude literals in the DEDICOM tensor factorization model used in this paper. The RESCAL model adds literals as $predicate - value$ pairs to a separate matrix [5,11,2]. A joint factorization is then made by considering this matrix as an additional constraint to the factorization algorithm. This approach has two drawbacks. First, it brought an additional computational complexity during the factorization due to the high dimensionality of the matrix. Second, processing the data into a predicate-value pairs is not efficient because of the possible number of comparisons especially in a MLOD environment where literals are mostly language dependent [2]. The other reason that motivate us to use entities' textual descriptions as an additional source of evidence is to solve some of the problems resulting from the open world assumption of the MLOD [12]. LDA is used to project entities textual description into a language independent space [35]. Finally, belief aggregation strategy is used to combine similarity analysis results computed from entities structural neighborhood and textual description into a global similarity score.

The main contributions of this paper are summarized as follows:

- We have adapted the DEDICOM tensor factorization model and showed for the first time its application to detect co-referent entity pairs even with a better performance to RESCAL in a MLOD context
- We have adopted a linear opinion pool, an approach used to aggregate probabilistic expert opinion to combine the structural and textual similarity of entities.
- We have proposed a language independent approach with an initial experiment that demonstrates a comparable result with language dependent approaches to compute textual similarity of entities across languages.
- We have proposed a mechanism to combine reasoning based approaches with SRL techniques based on higher order tensor factorization and provide an experiment that demonstrate a better result than using one of them alone.
- We have proposed an approach that enhance the performance of tensor factorization approaches by introducing predicate informativeness weight into the factorization algorithm.

The rest of this article is organized as follows: in section two, we briefly review related works. In section three, a detailed description about the proposed approach is made. Evaluation of the proposed approach is presented in section four followed by conclusion and future work in section five.

2. RELATED WORKS

Despite the fact that the RDF data model describes resources in a natural language independent manner, the process of creating automatic instance level links between data sets represented in different natural languages cannot be language independent due to the presence of language dependent textual descriptions (i.e. a very useful source of information to detect co-referent entity pairs) attached to resources and the high degree of heterogeneity in a MLOD environment[13].

Some authors apply statistical machine translation (SMT) to convert the cross-lingual entity co-reference resolution problem into a monolingual one which can then be solved by state of the art monolingual entity co-reference resolution systems. For instance, an approach is designed by translating such language specific descriptions of resources in[14]. The problem of this approach is that it can only be used for resources containing textual information. However,there are many resources in the LOD cloud that have no textual descriptions. Translating linguistic descriptions of resources through a SMT system is not easy due to problems related to translating ontological terms. One of the difficulties commonly observed in ontology translation is that the granularity level of some concepts may differ from culture to culture so this may result with incorrect translations[15]. Others believe the MLOD can be achieved by adding a layer of lexical resource over the existing LOD. This lexical resources enrich LOD datasets with lexical information required to establish instance level links in a MLOD environment. Such efforts includes EuroWordNet[1], BabelNet[16] and the various WordNets that have been translated in different natural languages such as GlobalWordNet[2] and meaning[3]. The problem of these approaches is that the number of natural languages covered by those lexical resources are few in number as compared to the number of natural languages joining the LOD cloud and these lexical resources contains domain-independent concepts and not cover lexical information on specific domains [17].

The proposed technique in this paper is closer to [5]. The authors introduced RESCAL to the problem under consideration in [5]. Following this, RESCAL is used to web scale entity co-reference resolution tasks in [11,2]. These experiments have showed the potential of RESCAL to find co-referent entities in the LOD environment but not its potential in a MLOD setting, a very different context. Our approach is based on the factorization of three-way DEDICOM thereby we introduce predicate quality weight as an additional constraint. Moreover, we also use reasoning over RDFS and OWL semantics to reduce the MLOD heterogeneity. Thus, this paper showed for the first time how much DEDICOM can be used to the task of entity co-reference resolution even with better performances than RESCAL.

3. ENTITY CO-REFERENCE RESOLUTION IN A MLOD CONTEXT

An entity co-reference resolution system which we named as ECRMLOD that considers the noise and heterogeneities of the MLOD environment is proposed. We combined SRL with reasoning based techniques to model the structural context of entities. LDA is used to model entities textual context into a language independent space. Finally, a belief aggregation technique is applied to combine similarity results from the structural and textual context. In the remaining subsections, a detailed descriptions about the proposed approach is made.

3.1 Structural Context

[1]http://www.illc.uva.nl/EuroWordNet/
[2]http://www.globalwordnet.org/
[3]http://adimen.si.ehu.es/wei4/doc/mcr/meaning.html

In the MLOD cloud a large number of structured knowledge from various domains is published in the form of Resource Description Framework (RDF) triples. The relationships of this network are important assets to detect co-referent entity pairs [5].

Definition 1 (Structural Context)..

A set of triples of the form $G = \{(s, p, o)\}$ defines the structural context over a domain D, where $s \in \{V\}$ is a subject entity, $p \in \{V\}$ denotes a relation type and $o \in \{V\}$ is an object entity. The vertices V are all the entities in D where entities in this context include all resources and classes. $P \subseteq V \times V$ is a set of all the relation types in D. $E \subseteq V \times V \times P$ is a set of known facts about entities in V.

There are factors that worsen the degree of structural heterogeneity and noise in a MLOD environment. First, some MLOD sources are obtained through an automatic extraction process so that noise could be introduced either from the process or from the nature of the data source itself. Second, there are many language dependent challenges. To mention some of these challenges: an average increase in schema heterogeneity is observed when the language diversity increases; there are also predicates that don't have meaning across languages. Furthermore, some predicates in the source language are not mapped to the target language, so identifying matchable properties in different natural languages is difficult. The misuse of ontological terms across language specific data sources is also another challenge. For example, the predicate *education* is used to refer to the education level in French DBpedia and education place in English DBpedia. The *BirthPlace* sometimes refers to the country of birth in some languages whereas it refers to the place of the birth in other languages. Third, resources among intra-language data sources share many things in common even if they are not co-referent as compared to resources among inter-language data sources.

All heterogeneities and noises discussed above affect conventional SRL techniques. Thus, we used existing explicit equivalent relationships and reasoned over them to merge syntactically distinct subject, predicate and object references into a common representation. Finally, we took into account the heterogeneity in the informativeness of the predicates through a weighting scheme. We quantify the informativeness of a predicate based on the diversity of the values it can take.

3.1.1 Reducing Heterogeneities

In a MLOD environment, the probability of seeing different syntactical representation among entity references is high even if they refer to the same entity. In addition to the effect on performance, this implies a scalability issue as the distinct syntactic representation increases the number of entities and relations.

To merge syntactically distinct URIs referring to the same entity into a common representation, the first construct used in this work is the semantics of $owl\colon sameAs$. $owl\colon sameAs$ indicates that all the URIs linked with this property, in the form of$(s, owl\colon sameAs, o)$, have the same identity. For instance,

$$(s1, p1, o1), (s2, p2, o2)$$

$$\underline{(s1, owl\colon sameAs, s2)}$$

$$\Rightarrow (s2, p1, o1) \vee (s1, p2, o2$$

Thus, the two distinct URI identifiers s1 and s2 can be merged into one representation.

The second construct which was used is the transitive property of $owl\colon sameAs$. For example,

$$(s1, p1, o1), (s2, p2, o2), (s3, p3, o3)$$

$$\underline{(s1, owl : sameAs, s2), (s2, owl : sameAs, s3)}$$

$$\Rightarrow (s3, p1, o1) \vee (s3, p2, o2) \vee (s2, p1, o1) \vee (s1, p2, o2)$$

This means the three distinct URI identifiers s1,s2 and s3 can be canonicalized into one. More equivalent relationships are also discovered by using the semantics provided by domain ontologies such as $owl\colon InverseFunctionalProperty$ and $owl\colon FunctionalProperty$.

3.1.2 Predicate informativeness Weight

We can classify the predicates based on their frequency of occurrence. For the co-reference resolution task, the very frequent predicates are not informative and can even be considered as potentially misleading since many entities share frequent predicates even if they are not co-referent. Also, rare predicates are not likely to help identify co-referent entities. Thus, to reduce the computational complexity, we removed both the frequent and the rare predicates.

Informative predicates lay below the most frequent predicates and above the rare predicates. These predicates are also not equally informative. Thus, assigning less weight for less-informative predicates and more weight for more-informative predicates can increase the performance of entity co-reference detection. The diversity of values a predicate can take is used to assign informativeness weight for predicates as shown in equation (1). The informativeness of a predicate k, referred to as Q(k), is computed by:

$$Q(k) = \frac{DO}{NT} \tag{1}$$

DO refers to distinct objects of k, NT to the total number of triples that use k.

We may wrongly consider two entities with syntactically distinct identifiers as two different entities though they are co-referent identifiers of one another. Thus, equation(1) assigns higher weight for predicates taking many co-referent object with distinct syntactical representation. This problem is referred as predicate weight overestimation in [18]. The problem needs special care in a MLOD environment due to the presence of syntactically distinct representation of co-referent objects. In order to reduce the predicate quality weight bias due to syntactical heterogeneity of MLOD, we compute predicate weight among intra-language data sources. For instance, for a predicate k, n quality weights$(k_{(1)}, k_{(2)}, ..., k$ are computed, where n is the number of languages(i.e. in our experiments 2) that predicate k is used. The final quality weight is then determined as the minimum quality weight $\min(k_{(1)}, k_{(2)}, ..., k_{(n)})$.

3.1.3 Tensor Factorization

We used a three-way DEDICOM tensor factorization model due to the following reasons [19,20,21]. First,the structure of DEDICOM model well fits to the properties of our data as the first two modes of the tensor are symmetric. This due to

the fact that entities appearing in the subject position also appears in the object position in any other time. Second, DEDICOM can solve the relaxed constraint of RESCAL[22]. Third, DEDICOM can compute a single latent space for entities appearing both in the first and second mode so that it has relational learning ability.

Given an entity i which appears in the subject position, an entity j in the object position, and a predicate k, the entries of the tensor X_{ijk} will be $1 + Q(k)$ if the k^{th} predicate exists between the i^{th} and j^{th} entity while 0 if the k^{th} predicate does not exists between the i^{th} and j^{th} entity, where $Q(k)$ is the quality weight attached to the k^{th} predicate. If the structural context in definition 1 has K unique predicates, then we can construct an adjacency matrix $\mathbf{X}_k$ for each predicate type, k = 1 . . . K, and store them as an array $\mathbf{X} \in \mathbb{R}^{I \times I \times K}$. DEDICOM factorizes the tensor $\mathbf{X} \in \mathbb{R}^{I \times I \times K}$ as follows;

$$\mathbf{X}_k = AD_kRD_kA^T, for k = 1, ..., K \quad (2)$$

Where $\mathbf{X}_k$ is the k^{th} adjacency matrix in $\mathbf{X}$, $A \in \mathbb{R}^{I \times r}$(rows of A represent entities in the domain, columns of A correspond to the latent components), $R \in \mathbb{R}^{r \times r}$, and $D_k \in \mathbb{R}^{r \times r}$ is a diagonal matrix that gives the weights of the columns of A for each predicate.

The decomposition of DEDICOM is done by minimizing the objective function in equation(3).

$$\min_{A,R,D} \sum_{k=1}^{K} \parallel X_k - (AD_kRD_kA^T) \parallel_F^2 \quad (3)$$

We adapted the ASALSAN (Alternating Simultaneous Approximation, Least Squares, and Newton) algorithm developed for computing a three-way DEDICOM model in a large and sparse data scenario in [22] to estimate A, R and D.

3.1.4 Structural Similarity

Rows of A correspond to entities in the domain whereas the columns can be considered the discovered latent variables. To explain the process, we take the example of figure-2. D indicate the strength of association of latent variables to the different predicates. From D, we can infer the proportion of the 3 latent variables across the 4 predicates (i.e. how much each predicate contributes to discover each latent variable). For instance, the first latent variable (i.e. lv_1) is composed of 0.2 of predicate 1 (p1), 0.3 of predicate 2 (P2), 0.2 of predicate 3 (P3) and 0.3 of predicate 4 (P4). On the other hand, we have a vector which shows the quality of each predicate (i.e. P1=0.6,P2=0.3,P3=0.2,P4=0.7) computed in equation(1). Thus, we can indirectly introduce quality to latent variable,Qlv_i, by equation(4).

$$Qlv_i = lv_i \times c_{p_1} \times Q_{p_1} + lv_i \times c_{p_2} \times Q_{p_2} +, ..., + lv_i \times c_{p_K} \times Q_{pK} \quad (4)$$

where lv_i refer the i^th latent variable, c_{p_K} the proportion of the i^th latent variable in the k^th predicate and Q_{pK} the quality attached to the k^th predicate.

We normalize the quality of latent variables to be in a range of 0 and 1 through a min-max normalization technique given in equation(5)[55]. Suppose that mi_{Qlv} and ma_{Qlv} respectively are the minimum and maximum quality weights of latent variables discovered in equation(4). Min-max normalization assigns a new quality weight,$Qlv_i^{'}$ to a latent variable

A=Nxr

Entity	LV_1	LV_2	LV_3
A1	0.1	0.2	0.03
A2	0.2	0.1	0.3
A3	0.4	0	0.1
A4	0.3	0.1	0.3
A5	0.001	0.04	0.6
A6	0.12	0.23	0.5
A7	0.45	0.12	0.12

D=Kxr

Predicate	LV_1	LV_2	LV_3
P1	0.2	0.1	0.4
P2	0.3	0.6	0.2
P3	0.3	0.2	0.1
P4	0.2	0.1	0.3

Predicate Weight
P1=0.6
P2=0.3
P3=0.2
P4=0.7

Figure 1: A and D in DEDICOM

value, Qlv_i in the range $[nmi_{Qlv}, nma_{Qlv}]$.

$$Qlv_i^{'} = \frac{Qlv_i - mi_{Qlv}}{ma_{Qlv} - mi_{Qlv}}($$

nma_{Qlv}-nmi_{Qlv})+new nmi_{Qlv} (5)

Where $nmin_{Qlv}$ and new max_{Qlv} respectively are the minimum and maximum values in the new range(i.e. in our experiments the new maximum value is 1 while the new minimum value is 0).

Assuming there are a total of N entities in the domain(i.e. total number of rows in a factor matrix A). Entities information is represented by r-dimensional vectors $f_1, ..., f_N$. The vector $f_i = (f_{i,1}, ..., f_{i,r})$ represent an entity i, where $f_{i,k}$ represents entity i's participation to the k^{th} discovered latent variable. With the assumption that co-referent entities are more structurally similar than non-co-referent ones, we used entities similarity score to decide whether they are co-referent or not. The similarity between entity i and j is computed as follows:

$$D(i,j) = \frac{\sqrt{\sum_{k=1}^{k} Qlv_k(i_k - j_k)^2}}{k} \quad (6)$$

where Qlv_k is the computed quality weight for the k^{th} latent variable.

Lastly, for each entity i, its structural similarity with the other N entities is computed using equation(6). This similarity score is used to find possible candidate for co-referent entities for an entity i. Thus, for an entity i, n entities where $\mathtt{n} \leq N$ having higher similarity score to i are considered as candidates.

3.2 Textual Context

Publishers using two different languages most probably have different knowledge of entities or they may want to see entities from possibly orthogonal point of views. It is natural for them to provide more detailed information for resources in their own native language. This contributes to the unbalance in the amount of data attached to resources across language specific LOD data sources. More importantly, two data sources may provide complementary information about resources [23]. For instance, in the French DBpedia, there is information about only the name of the capital city of Ethiopia whereas in YAGO (in English), information about location(i.e. altitude and longitude), neighboring regions, temperature conditions and demographic characteristics of the population living in the city can be accessed.

Since, majority of the digital information about entities is expressed in natural language text, one can use this text to

supplement the LOD cloud [24,25]. However, using this information requires an effective entity linking technique (i.e. the task of automatically extracting and assigning a text segment to a LOD entity) or links that connect a LOD entity to a text segment. There exists plenty of techniques to identify mentions of LOD entities in the web of documents [26,27,28,29,30,31]. Moreover, from our own observation, most of the LOD entities have links to descriptions in the form of web pages. Therefore, using unstructured textual data to describe LOD entities for the task of entity co-reference resolution is a feasible assumption. The algorithm which is used to integrate textual information from unstructured sources is presented in algorithm-1. The algorithm is efficient as it can be done off-line. We begin by defining the terminologies of the algorithm.

A LOD entity is characterized by a tuples $e = (uri, plink, dtext)$ where x_{uri} is a unique identifier for an entity e, x_{plink} is a value for a predicate that link entity e to the corresponding web page if any, x_{dtext} is a text segment which is a sequence of words $w_0, w_1, ..., w_i$ extracted from x_{uri} or informative attribute values. For an entity *Ethiopia* from DBpedia,

$$uriof(Ethiopia) = http://dbpedia.org/resource/Ethiopia,$$

$$plinkof(Ethiopia) = http://en.wikipedia.org/wiki/Ethiopia$$

which is the value of $foaf: primaryTopic$ predicate and dtext of (*Ethiopia*) can be the literal "Ethiopia" which is the value for $rdfs: label$ predicate or any other informative text. we use dtext value associated with the entity as an input to extract a *text segment* that describe the entity from unstructured sources while plink value of the entity is used to extract a *web page* about the entity. Suppose the set of all possible entities, text segments, and web pages are denoted by E, T and P respectively.

Integrating textual context for an entity $e \in E$ is the process of assigning $p \in P$ if plink(e) exists or extracting and assigning a text segment $t \in T$ to e. For an entity e, the web page $p \in P$ or a text segment $t \in T$ describing e are considered textual context of e. Textual context of e is denoted by $TexCon(e)$.

Algorithm 1:LOD entities textual information integration

Input: Entities in the domain, E
Initialize: $map \leftarrow \{\}$
1:for $e \in E$
2: if $plink(e)$ is true then
3: extract a web page $p \in P$
4: $map.put(e, p)$
5: else
6: extract a text segment $t \in T$
7: $map.put(e, t)$
8: end if
9:Return map
10:end for

3.2.1 Similarity of Textual Context

Given any two LOD entities i and j, their similarity is computed from the similarity of their contexts. However, textual context of entities are language dependent. Thus, we need a mechanism that compares textual descriptions of entities across languages. It is neither easy to build quality translation resources nor do parallel corpora exist for many language pairs and domains [32]. Due to this, probabilistic topic models that only require comparable corpora have gained much attention [33]. These models have been effectively used to find similar words across languages in [38] and cross-lingual information retrieval tasks[34]. In this research, Latent Dirichlet Allocation (LDA), a common probabilistic topic model is used to project the textual context of entities into a language independent latent topic space. The main reasons to prefer LDA over other topic models particularly Probabilistic Latent Semantic Analysis(i.e. a potential candidate for LDA) are the following[37, 29,35]; Probabilistic Latent Semantic Analysis(PLSA) has many parameters so it is more complex than LDA. These parameters also bring the existence of many local maxima which leads to over fitting. PLSA is not a real generative model and we cannot compute probabilities for a new document. Moreover, the performance of LDA is consistently better than other probabilistic topic models [35].

Particularly, a variant of LDA used in this paper is Bilingual LDA(BILDA). In LDA, documents are represented as random mixtures over latent topics, whereas each topic is characterized by a distribution over words. BILDA only requires theme aligned comparable corpus. BILDA uses the same topic distribution θ_m but a different word distributions for each topic.The topic $Z^S_{m,n}$ in the source language is sampled as

$$Z^S_{m,n} \sim P(Z^S_n \mid \theta_m) \tag{7}$$

The topic $Z^T_{m,n}$ in the target language is drawn as

$$Z^T_{m,n} \sim P(Z^T_n \mid \theta_m) \tag{8}$$

. Each word ω^s_{mn} in the vocabulary of the source language is then generated from a multinomial distribution

$$\omega^s_{mn} \sim P(\omega^s_n \mid Z^s_{m,n}, \varphi^s) \tag{9}$$

whereas a word ω^T_{mn} in the target language is generated by;

$$\omega^T_{mn} \sim P(\omega^T_n \mid Z^T_{m,n}, \Phi^T) \tag{10}$$

To infer previously unseen document d, we compute the probability of sampling a topic k given a document d as shown in equation (12).

$$P(Z_k \mid d) = \theta_k = \frac{n^{(}k)_d + \alpha_k}{\sum_{k=1}^{k}(n^{(}k)_d + \alpha_k)} \tag{11}$$

Where $n^{(}k)_d$ is the total number of times topic k is assigned to document d.

The data set used to train the BILDA is collected from the English and French Wikipedia editions and the European parliament corpus [36]. From wikipedia, we remove tables, references, images and info-boxes. We dropped all articles from French Wikipedia that is not linked to an English article. A total of 84,580 Wikipedia articles; 42,290 articles for each edition is collected. From the European parliament corpus, 357 parallel English and French speeches are used. Linguistic preprocessing tasks such as punctuation mark, tag, stop word removal is done. Porter and Snowball stemmers are employed for stemming English and French texts respectively. Cross-lingual alignment is then established for Wikipedia articles collected from French and English Wikipedia by using the Wikipedia inter-lingual link whereas cross-lingual speech level alignment is made for the text taken from the European parliament proceeding.

Some parameters must be set while training the model. One of the inputs the model expects is the number of topics

i.e., k. Setting the values of k to a very small number of topics will result with broader topics which leads to a case where non-similar documents can be wrongly classified as similar. Giving very large value for k slows down the system during the estimation of the word distribution. In addition to our empirical evidence, there is a body of research that has shown the effect of topic size to the overall optimality of the model [35]. In this experiment, we have checked the model with 100,200,400 topic spaces and discovered a better result when the number of topics is 200. Therefore, we used the value 0.005 for parameterα, 0.01 for β. Our model is trained with Gibbs sampling with 1000 iterations. Lastly, for an entity $e \in E$, the textual context is extracted by algorithm 1. After linguistic preprocessing, the textual context of e (i.e. $(TexCon(e))$) is projected into a language independent topic space by the formula in equation(11). Let $LDATexCon(i)$ be the projected textual context space of an entity i, $LDATexCon(j)$ is the projected textual space of entity j, the textual similarity between i and j is computed by the cosine similarity metrics in equation(12).

$$COS(i,j) = \frac{LDATexCon(i).LDATexCon(j)}{LDATexCon(i)||LDATexCon(i)} \quad (12)$$

3.3 Aggregating Evidences

This time, for an entity e, two n-dimensional vectors showing the similarity score of e to n candidate entities are obtained from the two sources(i.e. Structural and Textual). Our hypothesis is aggregating the two similarity analysis can optimize the result as the two sources are complementary to each other. However, the selection of the most suitable aggregation method depends on the nature the problem[39,40]. Tensor coupling based approaches are obviously computationally complex and produce sub optimal result with the presence of missing values which is the case in our. Thus, we used belief aggregation methods in this context work[41,42,43,44,45].

A detailed characterization of pooling functions and the main factors to be taken into consideration to select a pooling function is summarized in [45]. The purpose of the aggregation task in this research is to combine two complementary evidences so as to reach into a single global evidence. The first source of evidence is a similarity score between entities computed from the structural context produced after the tensor factorization while the latter is the similarity score of entities obtained from their textual context. The basic assumption is that the aggregated evidence can tell us more information than using one of them alone to identify co-referent entities in LOD. However, the performance of each of the sources varies across entities. For some entities, the structural context is more important than the textual evidence whereas the reverse is true for a significant number of entities. Hence, the pooling function to be used should provide the freedom of event-wise independence. This is a case in which the collective probability of any event solely depends on the individual probabilities of that event [45]. Moreover, for a given entity, we expect the pooling function to produce an aggregated result that can detect the right co-referent partners in situations when at least both of the evidences is able to detect co-referent entity pairs. This leads to a scenario where the pooling function should provide the unanimity preservation axiom, a situation in which if all sources in the input probability distributions hold the same evidence; these evidences become the collective ones[45,40].

Linear opinion pool in-general and Carvalho and Larson's consensual linear opinion pool in particular satisfy all the requirements mentioned above. Therefore, a consensual linear opinion pool algorithm proposed in [41] is used to aggregate similarity results. Let us represent the similarity score of an entity e from the structural similarity score by n-dimensional vector f_s and textual similarity score by n-dimensional probability vector f_t. The two vectors are converted into a probability distribution vector by equation(13).

$$\frac{X_i}{\sum_{i=0}^{n}(X_i)} \quad (13)$$

Where X_i are row entries and n is the length of the candidate entities.

Lastly, we consider f_s and f_t as two experts? beliefs about the similarity of entities(i.e. two experts f_s and f_t give their beliefs on how much entity e is similar to the other n entities). We use a consensual linear opinion pool algorithm presented in algorithm-2 to aggregate the two beliefs. The weight that an expert s assigns to expert t?s opinion at a given time t is computed as,

$$p_{s,t}^{(t)} = \frac{\alpha_s^t}{\epsilon + D(f_s^{t-1}, f_t^{t-1})} \quad (14)$$

where α_s^t normalizes the weights so that they sum to one,and ϵ is a small, positive constant used to avoid division by zero, D is the distance between f_s and f_t which is computed as follows;

$$D(f_s, f_t) = \frac{\sqrt{\sum_{k=1}^{n}(f_{sk} - f_{tk})^2}}{n} \quad (15)$$

Algorithm 2:Consensual Linear Opinion Pool[41]
Input: Source s, (f_s) and surce t,(f_t)
Initialize: Recalibration constant ϵ
1:for $t = 1 to \infty$ **do**
2: $p_{s,t}^{(t)} = \frac{\alpha_s^t}{\epsilon + D(f_s^{t-1}, f_t^{t-1})}$
3: $f_s^{(t)} = (p_{s,t})^{(t)} f_t^{(t-1)}$
4 $f_t^{(t)} = (p_{s,t})^{(t)} f_s^{(t-1)})$
5: end for

4. EXPERIMENTAL EVALUATION

4.1 Dataset Description

The main challenge to evaluate our system is the absence of standard benchmark datasets that can be representative to the amount of heterogeneity mentioned in this work. The Ontology Alignment Evaluation Initiative (OAEI) doesn't provide enough benchmark datasets to evaluate systems that can cope with multilingualism. The same problem is also reported in [14]. In the literature, several real datasets are identified as an alternative when OAEI is not appropriate[14,46]. These datasets include DBpedia, New York Times, Freebase, RKB, SWAT2 and the entire LOD. From the LOD cloud, the existing $owl:sameAs$ links between resources can be used to find test dataset.

Consequently, the dataset for the experiment is extracted from the French and English DBpedia sub graph by introducing SPARQL query[4]. The existing bi-directional $owl:sameAs$ links are used to find ground truth dataset. We manually

[4]https://github.com/melkamu-beyene/Code-contribution

check for correctness of $owl:sameAs$ links. From the Person and Institution subclass, sample RDF triples belonging to the depth first transitive closure of predicates starting with subjects and objects are extracted. A total of 4509 facts about 3007 resources connected by 171 predicates are extracted. By reasoning over rdfs and owl constructs, the total number of unique entities is reduced from 3007 to 2249 while the number of predicates is also reduced from 171 to 74. From the experiment data, we identified 400 entity mentions from the French DBpedia sub graph that are known to be co-referent with the English DBpedia sub graph. We then compute predicate quality weight for the 74 predicates.

For 74 predicates, a 2249×2249 matrix is built. Majority of the matrix entries are zero. For instance, the matrix for the $rdf:type$ predicate has only 3200 non-zero entries. In order to make our system memory efficient, we used python's compressed sparse row format (CSR) data structure to construct the adjacency matrix. The non-zero entries of the matrix are also one plus the predicate weight. After the adjacency matrix is built for the 74 predicates, it is converted to a $2249 \times 2249 \times 74$ tensor. The tensor is decomposed into a lower dimensional space through DEDICOM factorization model with a rank of 100. Then the structural similarity between entities is computed by equation(7). This similarity score is used to identify potential candidate co-referent entities for each entity. For each entity, its structural similarity with the other 2249 entities is computed. In the second stage, for each entity, textual descriptions are extracted from both the French and English wikipedia. Then, the textual similarity with the other candidate entities is computed. Finally, the two similarity vectors are converted into a collective one.

4.2 Discussion of Results

We compare our system (i.e. ECRMLOD) with entity co-reference resolution systems that are developed based on RESCAL tensor factorization to demonstrate the performance improvement. The RESCAL factorization model is introduced for the first time in[5]. Then, the authors provide a large scale experiment by factorizing YAGO[5] ontology to demonstrate its performance in[11]. RELATIONAL is the first modification to the original RESCAL model in[2]. This system report a good experimental result on OAEI 2010 and 2011 benchmarking datasets.

As far as the researchers' knowledge is concerned, neither RESCAL nor RELATIONAL has never been evaluated in a MLOD context. Consequently, we implemented the two algorithms to show how much our system (i.e. ECRMLOD) has brought performance improvement. After the implementation, we evaluate RESCAL and RELATIONAL with multi-lingual datasets prepared in this paper. Many entity co-reference resolution systems try to discover equivalent resources across data source(ontologies) through the assumption that a given data source don't contain co-referent resources [52,53,54]. However, to show ECRMLOD's consistent performance in detecting co-referent resources in cases where resource's source ontology(data source) and resource's target ontology(data source) are in the same or different language, precision is calculated by dividing the number of correctly detected co-referent pairs to the total number of resource pairs discovered to be co-referent out of the 2249×2249 possible combinations while recall is calculated by dividing the number of correctly detected co-referent pairs to 400 (i.e. the total co-referent pairs used as a ground truth in this experiment).

[5]http://yago-knowledge.org

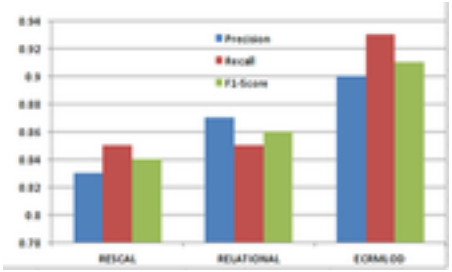

Figure 2: Performance Comparison

As shown in figure-2, ECRMLOD out performs conventional RESCAL and RELATIONAL in the three performance measures. This shows applying reasoning and predicate informativeness weight improve entity co-reference resolution system in a MLOD environment. Moreover, 5-7 percent F1-score performance increase is achieved by integrating unstructured textual information from unstructured sources. ECRMLOD has consistent performance among intra-language and inter-language data sources.

5. CONCLUSION AND FUTURE WORK

In this paper, an automatic mechanism of identifying co-referent entity pairs in the MLOD environment is proposed. We showed to what extent the existing SRL techniques can be enhanced when combined with reasoning techniques. Moreover, we used for the first time the DEDICOM tensor factorization model. We introduce relations weight into a tensor factorization and experimentally proved as DEDICOM is better than RESCAL to the task of entity co-reference resolution in a MLOD setting.

We have come-up with an effective technique that can solve syntactical heterogeneity and scalability problems related to the use of literals by introducing entities textual descriptions from unstructured sources as an additional source of knowledge. We provide an efficient and scalable algorithm to extract textual descriptions of LOD entities from the web of documents. We used a Bilingual LDA model to project textual description into a language independent latent space. A belief aggregation technique to aggregate similarities computed from the relationship of entities(tensor factorizations) and textual descriptions.

We also provide an experiment and a source code[6] that demonstrates a better performance improvement to the existing entity co-reference resolution systems in the literature. Currently, we are conducting a large scale experiment by taking multilingual datasets from the YAGO and DBpedia data sources.

6. REFERENCES

[1] Berners-Lee, T. (2010). Linked Data. W3C Design Issues, July 2006.

[2] de Assis Costa, G., and de Oliveira, J. M. P. A Relational Learning Approach for Collective Entity Resolution in the Web of Data.

[6]https://github.com/melkamu-beyene/Code-contribution

[3] Nickel, M. (2013). Tensor factorization for relational learning (Doctoral dissertation, lmu).

[4] Singla, P., and Domingos, P. (2006, December). Entity resolution with markov logic. In Data Mining, 2006. ICDM'06. Sixth International Conference on (pp. 572-582). IEEE.

[5] Nickel, M., Tresp, V., Kriegel, H. P. (2011). A three-way model for collective learning on multi-relational data. In Proceedings of the 28th international conference on machine learning (ICML-11) (pp. 809-816).

[6] S. Dzeroski. Relational data mining applications: an overview. In: Relational Data Mining. 2001, pp. 339?360. ???: http://dl.acm.org/citation.cfm?id=567240.

[7] A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Dis- covery and Data Mining. KDD ?08. New York, NY, USA: ACM, 2008, pp. 650?658. ????: 978-1-60558-193-4. ???: 10.1145/1401890.1401969. ???: http://doi.acm.org/ 10.1145/1401890.1401969.

[8] P. Singla and P. Domingos. Entity Resolution with Markov Logic. In: Proceedings of the Sixth International Conference on Data Mining. ICDM ?06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 572?582. ????: 0-7695-2701-9. ???: 10.1109/ICDM.2006.65. ???: http://dx.doi.org/10.1109/ICDM.2006.65.

[9] B. Taskar, M.-F. Wong, P. Abbeel, and D. Koller. Link Prediction in Relational Data. In: Advances in Neural Information Processing Systems. Ed. by S. Thrun, L. Saul, and B. Schölkopf. Vol. 16. NIPS?03. Cambridge, MA: MIT Press, 2004.

[10] L. Getoor and B. Taskar, eds. Introduction to statistical relational learning. MIT Press, 2007. ????: 0262072882. ???: http://www.cs.umd.edu/srl-book/.

[11] Nickel, M., Tresp, V., and Kriegel, H. P. (2012, April). Factorizing YAGO: scalable machine learning for linked data. In Proceedings of the 21st international conference on World Wide Web (pp. 271-280). ACM.

[12] Jiang, L., Wang, Y., Hoffart, J., Weikum, G. (2013, October). Crowdsourced entity markup. In Proceedings of the 1st International Conference on Crowdsourcing the Semantic Web-Volume 1030 (pp. 59-68). CEUR-WS. org.

[13] ?váb-Zamazal, O., Svátek, V. (2008). Analysing ontological structures through name pattern tracking. In Knowledge Engineering: Practice and Patterns (pp. 213-228). Springer Berlin Heidelberg.

[14] Lesnikova, T., David, J., and Euzenat, J. (2014). Interlinking English and Chinese RDF Data Sets Using Machine Translation. In Proc. 3rd ESWC workshop on Knowledge discovery and data mining meets linked open data (Know@ LOD), Hersounisos (GR) (Vol. 2013).,

[15] Montiel-Ponsoda, E., Gracia del Río, J., Aguado de Cea, G., Gómez-Pérez, A. (2011). Representing translations on the semantic web.

[16] Navigli, R., Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 193, 217-250.

[17] Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., McCrae, J. (2012). Challenges for the multilingual web of data. Web Semantics: Science, Services and Agents on the World Wide Web, 11, 63-71.

[18] Song, D. (2014). Towards a Linked Semantic Web: Precisely, Comprehensively and Scalably Linking Heterogeneous Data in the Semantic Web.

[19] Harshman, R. A. (1978, August). Models for analysis of asymmetrical relationships among N objects or stimuli. In First Joint Meeting of the Psychometric Society and the Society for Mathematical Psychology, McMaster University, Hamilton, Ontario.

[20] Kolda, T. G., Bader, B. W. (2009). Tensor decompositions and applications. SIAM review, 51(3), 455-500.

[21] Bro, R. (1998). Multi-way analysis in the food industry: models, algorithms, and applications (Doctoral dissertation, Københavns Universitet'Københavns Universitet', LUKKET: 2012 Det Biovidenskabelige Fakultet for Fødevarer, Veterinærmedicin og NaturressourcerFaculty of Life Sciences, LUKKET: 2012 Institut for FødevarevidenskabDepartment of Food Science, 2012 Institut for Fødevarevidenskab, 2012 Kvalitet og TeknologiDepartment of Food Science, Quality Technology).

[22] Bader, B. W., Harshman, R. A., Kolda, T. G. (2007, October). Temporal analysis of semantic graphs using ASALSAN. In Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on (pp. 33-42). IEEE.

[23] Suchanek, F. M., Abiteboul, S., Senellart, P. (2011). Paris: Probabilistic alignment of relations, instances, and schema. Proceedings of the VLDB Endowment, 5(3), 157-168.

[24] Scharffe, F., Fan, Z., Ferrara, A., Khrouf, H., Nikolov, A. (2011). Methods for automated dataset interlinking.

[25] Bretschneider, C., Oberkampf, H., Zillner, S. (2015). UIMA2LOD: Integrating UIMA Text Annotations into the Linked Open Data Cloud. In Knowledge Engineering and Semantic Web (pp. 16-31). Springer International Publishing.

[26] Chandel, A., Nagesh, P. C., Sarawagi, S. (2006, April). Efficient batch top-k search for dictionary-based entity recognition. In Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on (pp. 28-28). IEEE.

[27] Chaudhuri, S., Ganti, V., Xin, D. (2009, April). Exploiting web search to generate synonyms for entities. In Proceedings of the 18th international conference on World wide web (pp. 151-160). ACM.

[28] Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., Ciravegna, F. (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Web Semantics: science, services and agents on the World Wide Web, 4(1), 14-28.

[29] Wang, W., Xiao, C., Lin, X., Zhang, C. (2009, June). Efficient approximate entity extraction with edit distance constraints. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data (pp. 759-770). ACM.

[30] Bretschneider, C., Oberkampf, H., Zillner, S. (2015). UIMA2LOD: Integrating UIMA Text Annotations into the Linked Open Data Cloud. In Knowledge Engineering and Semantic Web (pp. 16-31). Springer International Publishing.

[31] Beheshti, S. M. R., Venugopal, S., Ryu, S. H., Benatallah, B., and Wang, W. (2013). Big data and cross-document coreference resolution: Current state and future opportunities. arXiv preprint arXiv:1311.3987.

[32] Vuli?, I., De Smet, W., Tang, J., and Moens, M. F. (2015). Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. Information Processing and Management, 51(1), 111-147.

[33] Zhu, Z., Li, M., Chen, L., Yang, Z. (2013, August). Building Comparable Corpora Based on Bilingual LDA Model. In ACL (2) (pp. 278-282).

[34] Roller, S., and Schulteim Walde, S. (2013). A multimodal LDA model integrating textual, cognitive and visual modalities. In Proceedings of the 2013 conference on empirical methods in natural language processing (EMNLP) (pp. 1146?1157).

[35] Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., McCallum, A. (2009, August). Polylingual topic models. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2 (pp. 880-889). Association for Computational Linguistics.

[36] Koehn, P. (2005, September). Europarl: A parallel corpus for statistical machine translation. In MT summit (Vol. 5, pp. 79-86).

[37] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

[38] Lu, Y., Mei, Q., and Zhai, C. (2011). Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA. Information Retrieval, 14(2), 178?203.

[39] Croft, W. B. (2000). Combining approaches to information retrieval. In Advances in information retrieval (pp. 1-36). Springer US.

[40] Allard, D., Comunian, A., and Renard, P. (2012). Probability aggregation methods in geoscience. Mathematical Geosciences, 44(5), 545-581.

[41] Carvalho, A., and Larson, K. (2013, August). A consensual linear opinion pool. In Proceedings of the Twenty-Third international joint conference on Artificial Intelligence (pp. 2518-2524). AAAI Press.

[42] Rufo, M. J., Martín, J., and Pérez, C. J. (2012). Log-linear pool to combine prior distributions: A suggestion for a calibration-based approach. Bayesian Analysis, 7(2), 411-438.

[43] Heskes, T. (1998). Selecting weighting factors in logarithmic opinion pools. Advances in neural information processing systems, 266-272.

[44] Clemen, R. T., and Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. Risk analysis, 19(2), 187-203.

[45] Franz D., Christian L.(2014), Probabilistic Opinion Pooling,

[46] Dezhao Song, Scalable and Domain-Independent Entity Coreference: Establishing High Quality Data Linkages Across Heterogeneous Data Sources

[47] Lesnikova, T., David, J., and Euzenat, J. (2014). Interlinking English and Chinese RDF Data Sets Using Machine Translation. In Proc. 3rd ESWC workshop on Knowledge discovery and data mining meets linked open data (Know@ LOD), Hersounisos (GR) (Vol. 2013).

[48] D. Dey, V. S. Mookerjee, and D. Liu, ?Efficient techniques for online record linkage,? IEEE Trans. Knowl. Data Eng., vol. 23, no. 3, pp. 373?387, 2011.

[49] W. Hu and Y. Qu, ?Falcon-ao: A practical ontology matching system,? Journal of Web Semantics, vol. 6, no. 3, pp. 237?239, 2008.

[50] J. Li, J. Tang, Y. Li, and Q. Luo, ?RiMOM: A dynamic multistrategy ontology alignment framework,? Knowledge and Data Engineering, IEEE Transactions on, vol. 21, no. 8, pp. 1218?1232, aug. 2009.

[51] N. Aswani, K. Bontcheva, and H. Cunningham, ?Mining information for instance unification,? in The 5th International Semantic Web Conference (ISWC), 2006, pp. 329?342.

[52] C. Matuszek, J. Cabral, M. Witbrock, and J. Deoliveira. An introduction to the syntax and content of Cyc. In Proc. AAAI Spring Symposium, 2006.

[53] I. Niles and A. Pease. Towards a standard upper ontology. In Proc. FOIS, pages 2?9, 2001.

[54] F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO:A core of semantic knowledge. Unifying WordNet and Wikipedia. In Proc. WWW, pages 697?706, 2007.

[55] Han, J., Kamber, M., Pei, J. (2011). Data mining: concepts and techniques. Elsevier.

[56] R. A. Harshman and M. E. Lundy, Three-way DEDICOM: Analyzing multiple matrices of asymmetric relationships. Paper presented at the Annual Meeting of the North American Psychometric Society, 1992.